# A note on the Bayesian regret of Thompson Sampling with an arbitrary prior

Sébastien Bubeck, Che-Yu Liu
Department of Operations Research and Financial Engineering,
Princeton University
sbubeck@princeton.edu, cheliu@princeton.edu

April 23, 2013

**Abstract**

We consider the stochastic multi-armed bandit problem with a prior distribution on the reward distributions. We show that for any prior distribution, the Thompson Sampling strategy achieves a Bayesian regret bounded from above by $14\sqrt{nK}$. This result is unimprovable in the sense that there exists a prior distribution such that any algorithm has a Bayesian regret bounded from below by $\frac{1}{20}\sqrt{nK}$.

In this paper we are interested in the Bayesian multi-armed bandit problem which can be described as follows. Let $\pi_0$ be a known distribution over some set $\Theta$, and let $\theta$ be a random variable distributed according to $\pi_0$. For $i \in [K]$, let $(X_{i,s})_{s\geq 1}$ be identically distributed random variables taking values in $[0,1]$ and which are independent conditionally on $\theta$. Denote $\mu_i(\theta) := \mathbb{E}(X_{i,1}|\theta)$. Consider now an agent facing $K$ actions (or arms). At each time step $t = 1, \dots n$, the agent pulls an arm $I_t \in [K]$. The agent receives the reward $X_{i,s}$ when he pulls arm $i$ for the $s^{th}$ time. The arm selection is based only on past observed rewards and potentially on an external source of randomness. More formally, let $(U_s)_{s\geq 1}$ be an i.i.d. sequence of random variables uniformly distributed on $[0,1]$, and let $T_i(s) = \sum_{t=1}^{s} \mathbb{1}_{I_t = i}$, then $I_t$ is a random variable measurable with respect to $\sigma(I_1, X_{1,1}, \dots, I_{t-1}, X_{I_{t-1}, T_{I_{t-1}}(t-1)}, U_t)$. We measure the performance of the agent through the Bayesian regret defined as

$$R_n = \mathbb{E} \sum_{t=1}^{n} \left( \max_{i \in [K]} \mu_i(\theta) - \mu_{I_t}(\theta) \right),$$

where the expectation is taken with respect to the parameter $\theta$, the rewards $(X_{i,s})_{s\geq 1}$, and the external source of randomness $(U_s)_{s\geq 1}$.

The multi-armed bandit problem has a long history and we refer the interested reader to Bubeck and Cesa-Bianchi [2012] for a survey of this extensive literature. In this paper we are interested in studying the Thompson Sampling strategy which was proposed in the very

first paper on the multi-armed bandit problem Thompson [1933]. The strategy can be described very succinctly: let $\pi_t$ be the posterior distribution on $\theta$ given the history $H_t = (I_1, X_{1,1}, \dots, I_{t-1}, X_{I_{t-1}, T_{I_{t-1}}(t-1)})$ of the algorithm up to the beginning of round $t$. Then Thompson Sampling first draws a parameter $\theta_t$ from $\pi_t$ (independently from the past given $\pi_t$) and it pulls $I_t \in \mathrm{argmax}_{i \in [K]} \mu_i(\theta_t)$.

Recently there has been a surge of interest for this simple policy, mainly because of its flexibility to incorporate prior knowledge on the arms, see for example Chapelle and Li [2011]. For a long time the theoretical properties of Thompson Sampling remained elusive. The specific case of binary rewards with a Beta prior is now very well understood thanks to the papers Agrawal and Goyal [2012a], Kaufmann et al. [2012], Agrawal and Goyal [2012b]. In particular the last paper shows that in this specific setting the regret is bounded from above by $C\sqrt{nK \log n}$ for some numerical constant $C > 0$. This result was greatly generalized[1] by Russo and Roy [2013] who proved that in fact this is true for *any* prior distribution $\pi_0$. Precisely they show that Thompson Sampling always satisfies $R_n \leq 5\sqrt{nK \log n}$. Our main result is to show that the extraneous logarithmic factor in these bounds can be removed by using ideas reminiscent of the MOSS algorithm of Audibert and Bubeck [2009]. Precisely we prove the following theorem.

**Theorem 1** *For any prior distribution $\pi_0$ Thompson Sampling satisfies*

$$R_n \leq 14\sqrt{nK}.$$

Remark that the above result is unimprovable in the sense that there exist prior distributions $\pi_0$ such that for any algorithm one has $R_n \geq \frac{1}{20}\sqrt{nK}$ (see e.g. [Theorem 3.5, Bubeck and Cesa-Bianchi [2012]]). This theorem also implies an optimal rate of identification for the best arm, see Bubeck et al. [2009] for more details on this.

**Proof** We decompose the proof into three steps. We denote $i^*(\theta) \in \mathrm{argmax}_{i \in [K]} \mu_i(\theta)$, in particular one has $I_t = i^*(\theta_t)$.

**Step 1: rewriting of the Bayesian regret in terms of upper confidence bounds.** This step is given by [Proposition 1, Russo and Roy [2013]] which we reprove for sake of completness. Let $B_{i,t}$ be a random variable measurable with respect to $\sigma(H_t)$. Note that by definition $\theta_t$ and $\theta$ are identically distributed conditionally on $H_t$. This implies by the tower rule:

$$\mathbb{E} B_{i^*(\theta),t} = \mathbb{E} B_{i^*(\theta_t),t} = \mathbb{E} B_{I_t,t}.$$

Thus we obtain:

$$\mathbb{E}\left(\mu_{i^*(\theta)}(\theta) - \mu_{I_t}(\theta)\right) = \mathbb{E}\left(\mu_{i^*(\theta)}(\theta) - B_{i^*(\theta),t}\right) + \mathbb{E}\left(B_{I_t,t} - \mu_{I_t}(\theta)\right).$$

Inspired by the MOSS strategy of Audibert and Bubeck [2009] we will now take

$$B_{i,t} = \hat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{\log_+\left(\frac{n}{KT_i(t-1)}\right)}{T_i(t-1)}},$$

[1] Note however that the result of Agrawal and Goyal [2012b] applies to the *individual* regret (for $\theta$ fixed) while the result of Russo and Roy [2013] only applies to the integrated Bayesian regret.

where $\hat{\mu}_{i,s} = \frac{1}{s}\sum_{t=1}^{s} X_{i,t}$, and $\log_+(x) = \log(x)\mathbb{1}_{x\geq 1}$. In the following we denote $\delta_0 = 2\sqrt{\frac{K}{n}}$. From now on we work conditionally on $\theta$ and thus we drop all the dependency on $\theta$.

**Step 2: control of $\mathbb{E}\left(\mu_{i^*(\theta)}(\theta) - B_{i^*(\theta),t}|\theta\right)$.** By a simple integration of the deviations one has

$$\mathbb{E}\left(\mu_{i^*} - B_{i^*,t}\right) \leq \delta_0 + \int_{\delta_0}^{1} \mathbb{P}(\mu_{i^*} - B_{i^*,t} \geq u) du.$$

Next we extract the following inequality from Audibert and Bubeck [2010] (see p2683–2684), for any $i \in [K]$,

$$\mathbb{P}(\mu_i - B_{i,t} \geq u) \leq \frac{4K}{nu^2} \log\left(\sqrt{\frac{n}{K}} u\right) + \frac{1}{nu^2/K - 1}.$$

Now an elementary integration gives

$$\begin{aligned}\int_{\delta_0}^{1} \frac{4K}{nu^2} \log\left(\sqrt{\frac{n}{K}} u\right) du = \left[-\frac{4K}{nu} \log\left(e\sqrt{\frac{n}{K}} u\right)\right]_{\delta_0}^{1} &\leq \frac{4K}{n\delta_0} \log\left(e\sqrt{\frac{n}{K}}\delta_0\right) \\ &= 2(1+\log 2)\sqrt{\frac{K}{n}},\end{aligned}$$

and

$$\begin{aligned}\int_{\delta_0}^{1} \frac{1}{nu^2/K - 1} du = \left[-\frac{1}{2}\sqrt{\frac{K}{n}} \log\left(\frac{\sqrt{\frac{n}{K}}u + 1}{\sqrt{\frac{n}{K}}u - 1}\right)\right]_{\delta_0}^{1} &\leq \frac{1}{2}\sqrt{\frac{K}{n}} \log\left(\frac{\sqrt{\frac{n}{K}}\delta_0 + 1}{\sqrt{\frac{n}{K}}\delta_0 - 1}\right) \\ &= \frac{\log 3}{2}\sqrt{\frac{K}{n}}.\end{aligned}$$

Thus we proved: $\mathbb{E}\left(\mu_{i^*(\theta)}(\theta) - B_{i^*(\theta),t}|\theta\right) \leq \left(2 + 2(1+\log 2) + \frac{\log 3}{2}\right)\sqrt{\frac{K}{n}} \leq 6\sqrt{\frac{K}{n}}$.

**Step 3: control of $\sum_{t=1}^{n} \mathbb{E}\left(B_{I_t,t} - \mu_{I_t}(\theta)|\theta\right)$.** We start again by integrating the deviations:

$$\mathbb{E}\sum_{t=1}^{n}\left(B_{I_t,t} - \mu_{I_t}\right) \leq \delta_0 n + \int_{\delta_0}^{+\infty} \sum_{t=1}^{n} \mathbb{P}(B_{I_t,t} - \mu_{I_t} \geq u) du.$$

Next we use the following simple inequality:

$$\sum_{t=1}^{n} \mathbb{1}\{B_{I_t,t} - \mu_{I_t} \geq u\} \leq \sum_{s=1}^{n}\sum_{i=1}^{K} \mathbb{1}\left\{\hat{\mu}_{i,s} + \sqrt{\frac{\log_+\left(\frac{n}{Ks}\right)}{s}} - \mu_i \geq u\right\},$$

which implies

$$\sum_{t=1}^{n} \mathbb{P}(B_{I_t,t} - \mu_{I_t} \geq u) \leq \sum_{i=1}^{K}\sum_{s=1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} + \sqrt{\frac{\log_+\left(\frac{n}{Ks}\right)}{s}} - \mu_i \geq u\right).$$

Now for $u \geq \delta_0$ let $s(u) = \lceil 3\log\left(\frac{nu^2}{K}\right)/u^2 \rceil$ where $\lceil x \rceil$ is the smallest integer large than $x$. Let $c = 1 - \frac{1}{\sqrt{3}}$. It is is easy to see that one has:

$$\sum_{s=1}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} + \sqrt{\frac{\log_+\left(\frac{n}{Ks}\right)}{s}} - \mu_i \geq u\right) \leq \frac{3\log\left(\frac{nu^2}{K}\right)}{u^2} + \sum_{s=s(u)}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} - \mu_i \geq cu\right).$$

Using an integration already done in Step 2 we have

$$\int_{\delta_0}^{+\infty} \frac{3\log\left(\frac{nu^2}{K}\right)}{u^2} \leq 3(1+\log(2))\sqrt{\frac{n}{K}} \leq 5.1\sqrt{\frac{n}{K}}.$$

Next using Hoeffding's inequality and the fact that the rewards are in $[0,1]$ one has for $u \geq \delta_0$

$$\sum_{s=s(u)}^{n} \mathbb{P}\left(\hat{\mu}_{i,s} - \mu_i \geq cu\right) \leq \sum_{s=s(u)}^{n} \exp(-2sc^2u^2)\mathbb{1}_{u\leq 1/c} \leq \frac{\exp(-12c^2\log 2)}{1-\exp(-2c^2u^2)}\mathbb{1}_{u\leq 1/c}.$$

Now using that $1-\exp(-x) \geq x - x/2$ for $x \geq 0$ one obtains

$$\begin{aligned}
\int_{\delta_0}^{1/c} \frac{1}{1-\exp(-2c^2u^2)}du &= \int_{\delta_0}^{1/(2c)} \frac{1}{1-\exp(-2c^2u^2)}du + \int_{1/(2c)}^{1/c} \frac{1}{1-\exp(-2c^2u^2)}du \\
&\leq \int_{\delta_0}^{1/(2c)} \frac{1}{2c^2u^2 - 2c^4u^4}du + \frac{1}{2c(1-\exp(-1/2))} \\
&\leq \int_{\delta_0}^{1/(2c)} \frac{2}{3c^2u^2}du + \frac{1}{2c(1-\exp(-1/2))} \\
&= \frac{2}{3c^2\delta_0} - \frac{4}{3c} + \frac{1}{2c(1-\exp(-1/2))} \\
&\leq 1.9\sqrt{\frac{n}{K}}.
\end{aligned}$$

Putting the pieces together we proved

$$\mathbb{E}\sum_{t=1}^{n}\left(B_{I_t,t} - \mu_{I_t}\right) \leq 7.6\sqrt{nK},$$

which concludes the proof together with the results of Step 1 and Step 2. ■

# References


S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012a.

S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling, 2012b. arXiv:1209.3353.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*, 2009.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: an asymptotically optimal finite-time analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT)*, 2012.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling, 2013. arXiv:1301.2609.

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Bulletin of the American Mathematics Society*, 25:285–294, 1933.